Question:

What algorithm, machine learning, or AI approaches would you take to find anomalies in the duration of a span? And why do you think that approach is a good approach?

Answer:

Firstly I will use **K-means clustering**, then I will use **Isolation forest** followed by **One-Class SVM** for detecting anomalies in the duration time.

Why and How:

1. **K-means Clustering:**
    i. Initially, I will use K-means clustering to group the data points based on their features (duration time in this case).
    ii. By clustering the data, I aim to identify groups of spans with similar durations.
    iii. Anomalies could potentially be data points that do not belong to any cluster or belong to a small cluster with significantly fewer points.
2. **Isolation Forest:**
    i. Next, I will apply the Isolation Forest algorithm.
    ii. Isolation Forest is effective for anomaly detection because it isolates anomalies by randomly partitioning the data.
    iii. Anomalies are expected to be isolated in fewer partitions than normal data points, making them easier to detect.
3. **One-Class SVM:**
    i. Finally, I will use One-Class SVM, which is designed specifically for outlier detection.
    ii. One-Class SVM learns a boundary that separates normal data points from anomalies in a high-dimensional space.
    iii. It's particularly useful when you have only normal data for training and want to detect deviations from this normal behavior.